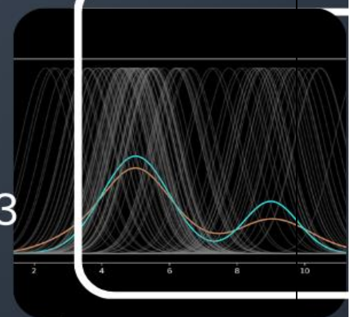
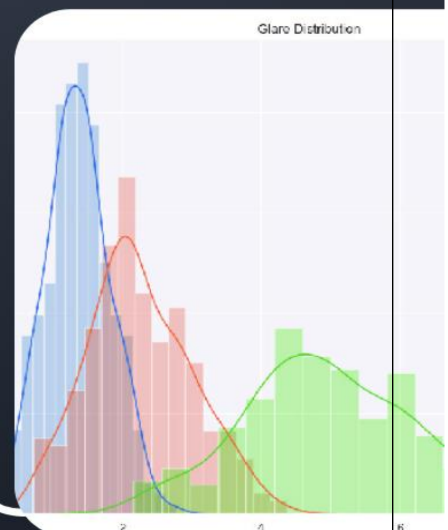
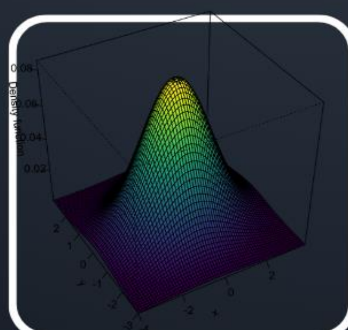
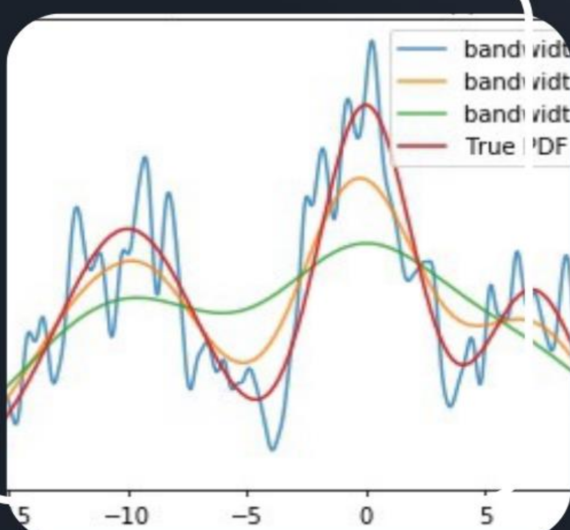
DENSITY ESTIMATION

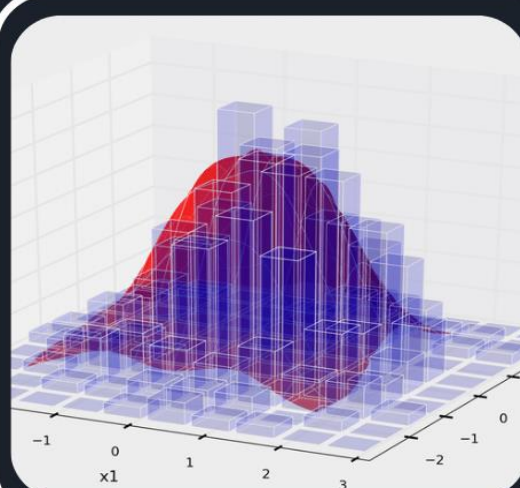
A report submitted as per the requirements of IS 3003
S15680

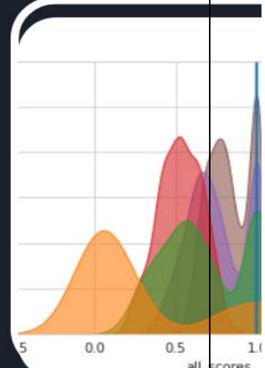


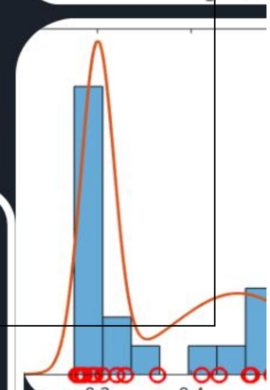












Abstract

Density estimation is a collection of methods for constructing an estimate of a probability density, as a function of an observed sample of data. Comprehensive overview of different density estimation methods will be given by this report focusing on univariate, bivariate and multivariate aspects.

Table of Contents

List of Figures

List of Tables

1.Introduction

1.1 Density Estimations

Density estimation is a collection of methods for constructing an estimate of a probability density, as a function of an observed sample of data.

2.Univariate Density Estimations

2.1 Histograms

Given class intervals of equal width h, the histogram density estimate based on a sample size n is

$\hat{f} = \frac{\mathrm{Vk}}{nh}$ h = class width.

Vk =number of points in the interval.

n = Sample size.

The questions arise in histogram are **how to decide the number of bins and boundaries** and **width of the class intervals** where undersmoothing narrow the bin with and over smoothing wide the bin width

Methods to decide boundaries of class intervals

1. Sturges'rule
2. Scott's normal reference Rule
3. Freedman-Diaconis rule

1. Sturges'rule

Optimal width of a class intervals = $\frac{\boldsymbol{R}}{\boldsymbol{1+log_2 n}}$: R = sample range.

This method is based on normal assumption and do not perform well for skewed distribution or multimodal distribution and trending to oversmooth

Example_Sturges'rule:

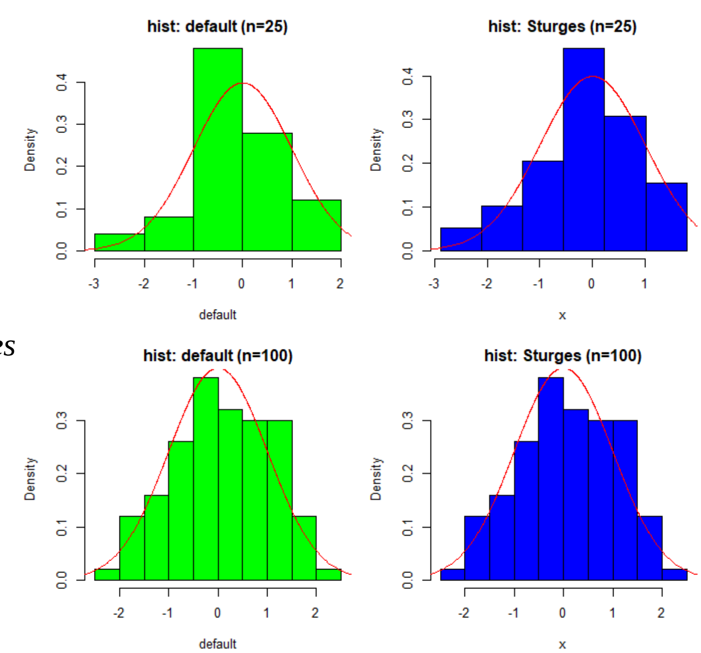


Figure 1 Comparison of default histogram and struges

2. Scott's Normal Reference Rule

Mean squared error (MSE) of a density estimator $\hat{f}$;

MSE($\hat{f}(x)$) =E[$\hat{f}(x) - f(x)$)]^2

To evaluate ˆf ISE is considered

ISE($\hat{f}(x)$) =$\int[f^{\hat{}}(x) - f(x))]$〗 ^2

MISE($\hat{f}(x)$) =E[ISE($\hat{f}(x)$)]

MISE=$\frac{1}{nh} + \frac{h^2}{12}\int f'(x)^2$ dx + O$\left[\frac{1}{n} + h^3\right]$

Optimal choice bin width = h_n* = $\left[\frac{6n}{f\prime(x)^2dx}\right]^{2}$

Scott's normal reference rule: $\hat{h}$ = 3.49$\hat{\sigma}$n $^{-1/3}$

Example_Density_estimation_for_Old_Faithful

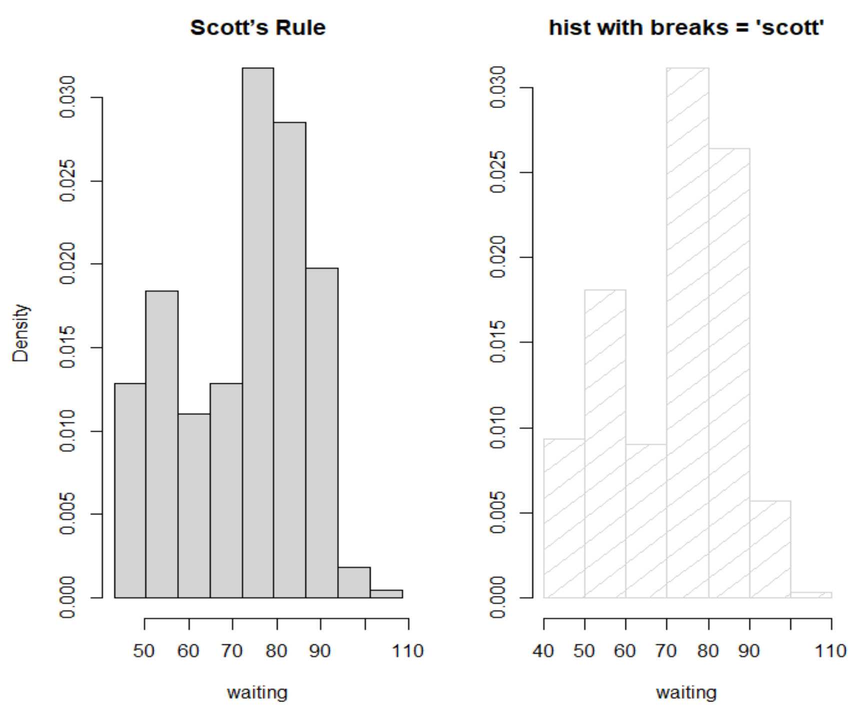


Figure 2 Comparison of histogram using Scott's rule and hist with breacks="Scott"

3. Freedman-Diaconis Rule

This is Less sensitive to outliers and have a trend to undersmooth.

Optimal choice of bin width: is $\hat{h} = 2(\mathrm{IQR})n^{-1/3}$

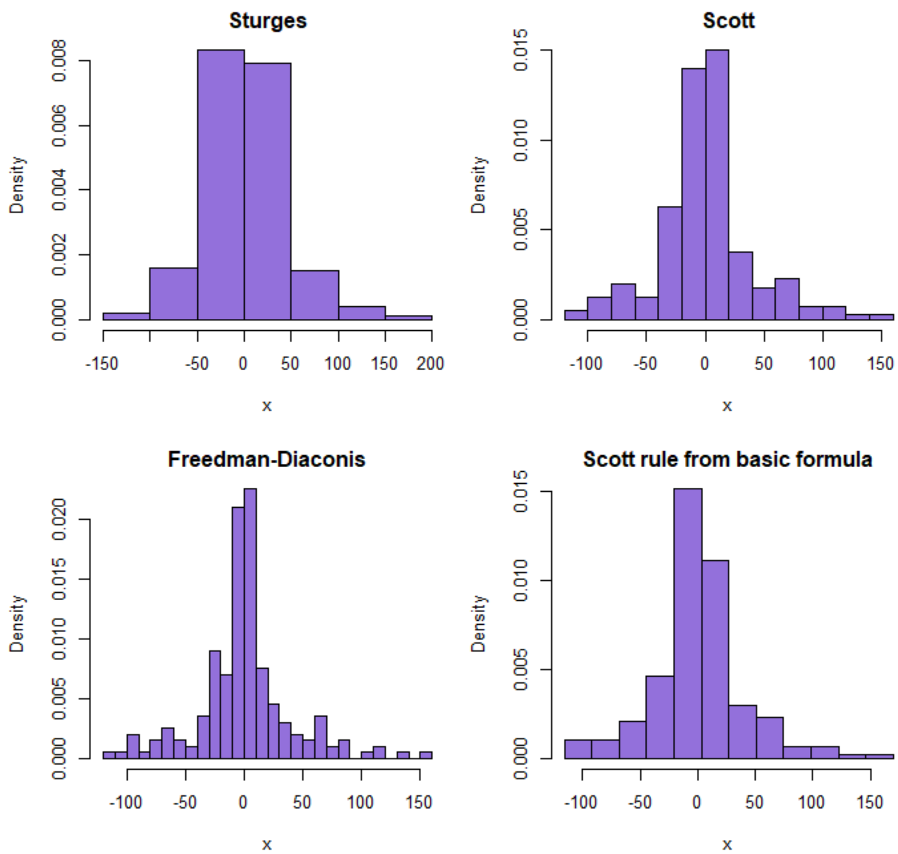


Figure 3 Comparison of Models for Determining class intervals in Histograms

2.2 Frequency Polygon Density Estimate

Histograms	Frequency polygon
• Not continuous • Represent the distribution dividing the range into intervals and counting observations in separate bins	• Continuous using linear interpolation • Provides a smooth representation

Table 1 Comparison of Histogram and Frequency polygon

For normal density, $h_n^{fp} = 2\left[\frac{49}{15\int f''(x)^2\,dx}\right]^{-\frac{1}{5}} * n^{-1/5}$

h_n^{fp} =2.15 $\sigma n^{-1/5}$

Example:A frequency polygon density estimate of the waiting time in the geyser data in MASS package.

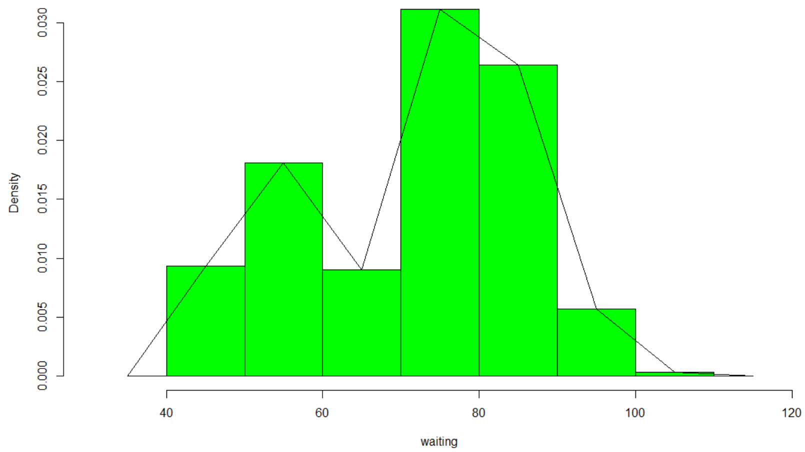


Figure 4 Frequency polygon estimate of Old Faithful waiting time density

2.3 The Average Shifted Histogram

Since the selection of "the optimal bin width," doesn't tell the location of the bin should be consider averaging several shifted histograms.

The average shifted histogram (ASH) estimate of density

$\boldsymbol{f}_{\boldsymbol{ASH}}\widehat{(\boldsymbol{x})} = \frac{\boldsymbol{1}}{\boldsymbol{m}}\sum_{\boldsymbol{j=1}}^{\boldsymbol{m}} \widehat{\boldsymbol{f}_{\boldsymbol{\iota}}(\boldsymbol{x})}$ the class boundaries for estimate $\boldsymbol{f}_{\boldsymbol{i}}(\boldsymbol{x})$ are shifted by h m from the boundaries

Optimal bin width: h^* =2.57 $\sigma n^{-1/5}$

Constructing an ASH density estimate of the waiting time in the geyser data, based on 20 shifted histograms .bin width (h) = 7.27037.

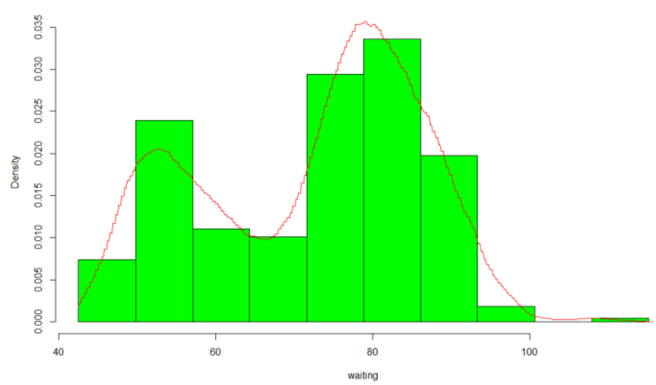


Figure 5ASH density estimates of Old Faithful waiting time density

2.4 Kernel Density Estimation

KDE generalizes histogram density estimate using a sample. e. If a histogram with bin width h is constructed from a sample X1,...,Xn, then a density estimate for a point x within the range of the data is;

$\widehat{f(x)} = \frac{1}{2hn} * k$ k =number of sample points in (x-h,x+h).

The density estimator(naïve density estimator)

$\widehat{f(x)} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}w\,(\frac{x-Xi}{h})$ w $=\frac{1}{2}I(|t| < 1)$weight of a function

KDE replaces w(t) with K(.) such that $\int_{-\alpha}^{\alpha} k(t) = 1$

$\widehat{f(x)} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(.)\,(\frac{x-Xi}{h})$

2.4.1 Kernel Density Estimates Using a Gaussian Kernel with Different Band Widths

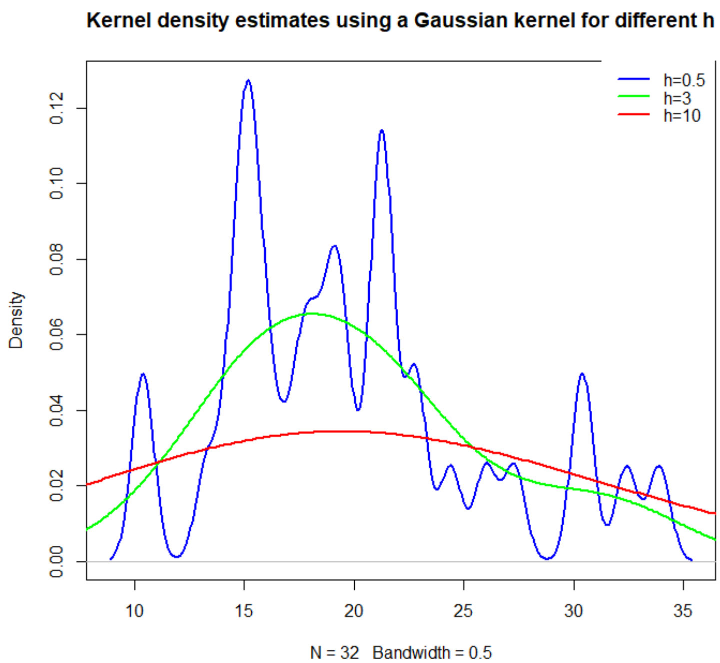
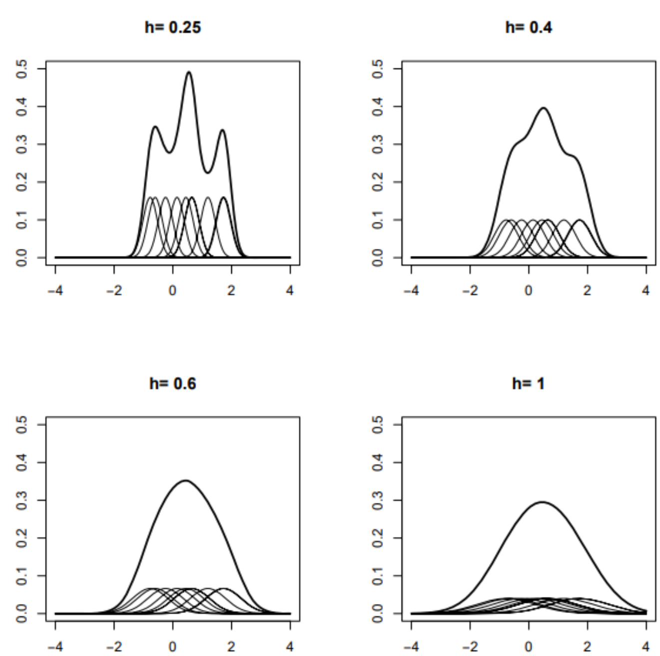


Figure 6 KDE Using a Gaussian Kernel with Different Band Widths



As the window width h decreases, the density estimate becomes rougher, and larger h corresponds to smoother density estimates

2.4.2 Kernel Functions that are Commonly Applied in Density Estimation

Kernel	K(t)	Support	σ_k^2	Efficiency
Gaussian	$\frac{1}{\sqrt{2\pi}}$ exp(-1/2*t^2)	R	1	1.0513
Epanechnikov	$\frac{3}{4}(1-t^2)$	|t|<1	1/5	1
Rectangular	1/2	|t|<1	1/3	1.0758
Triangular	1-|t|	|t|<1	1/6	1.0061
Biweight	15/16(1-t^2)^2	|t|<1	1/7	1.00610
Cosine	$\frac{\pi}{4} \cos \frac{\pi}{2} t$	R	1-8/π^2	1.0005

Table 2 Table kernel functions that are commonly applied in density estimation

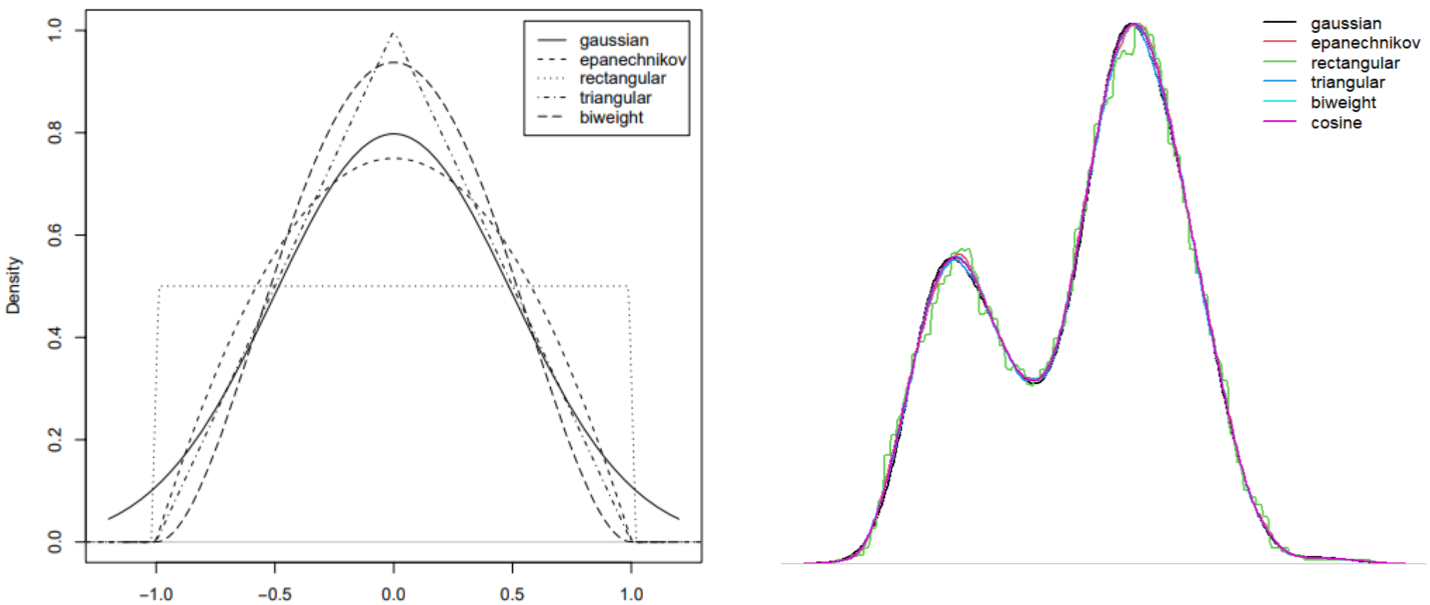


Figure 7 Different kernel functions that are commonly applied in Density Estimations

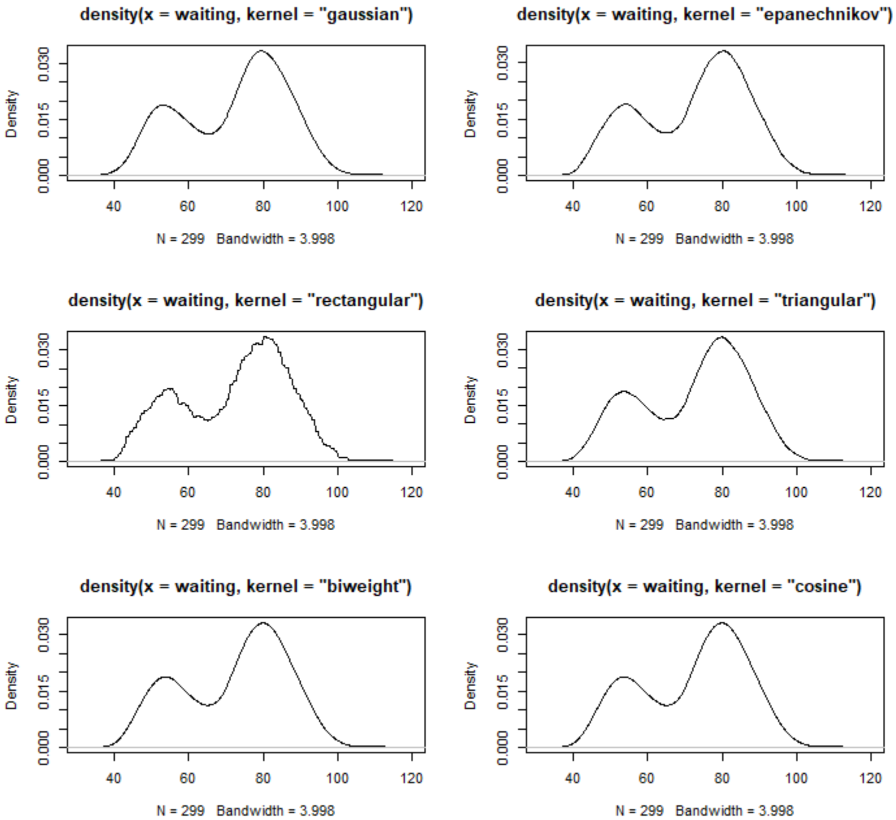
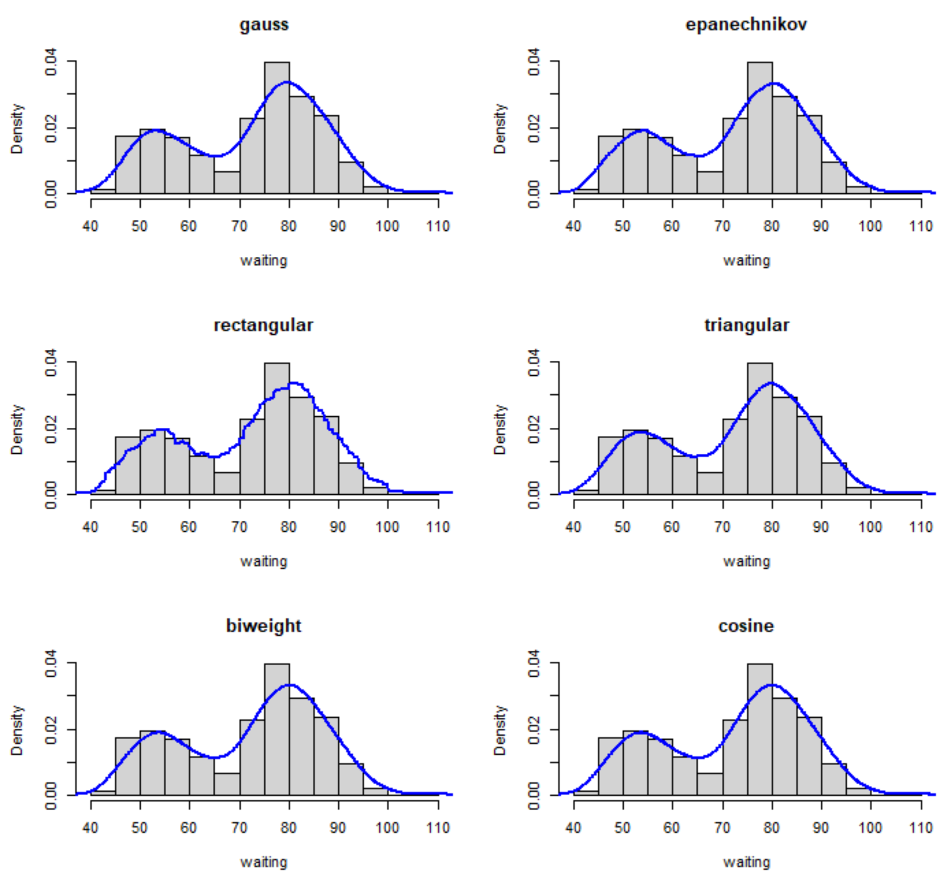


Figure 8 Different kernel functions that are commonly applied in Density Estimations using example



For a Gaussian kernel and a normal distribution, h=1.06$\sigma n^{-1/5}$ optimizes IMSE.If the density is not a unimodal then bandwit(h) tend to oversmooth.

The Silverman's "rule of thumb"

H=0.9min(s,IQR/1.34)$n^{-1/5}$ performs better for a wide range of distributions

2.5 Boundary Kernels

KDE estimates have larger errors near the boundaries of the support set of a density, or discontinuity points. KDE estimates tend to smooth the probability mass over the discontinuity points or boundary points

Example_(Exponential_density): boundary problem with an exponential density,compare the kernel estimate with the true density

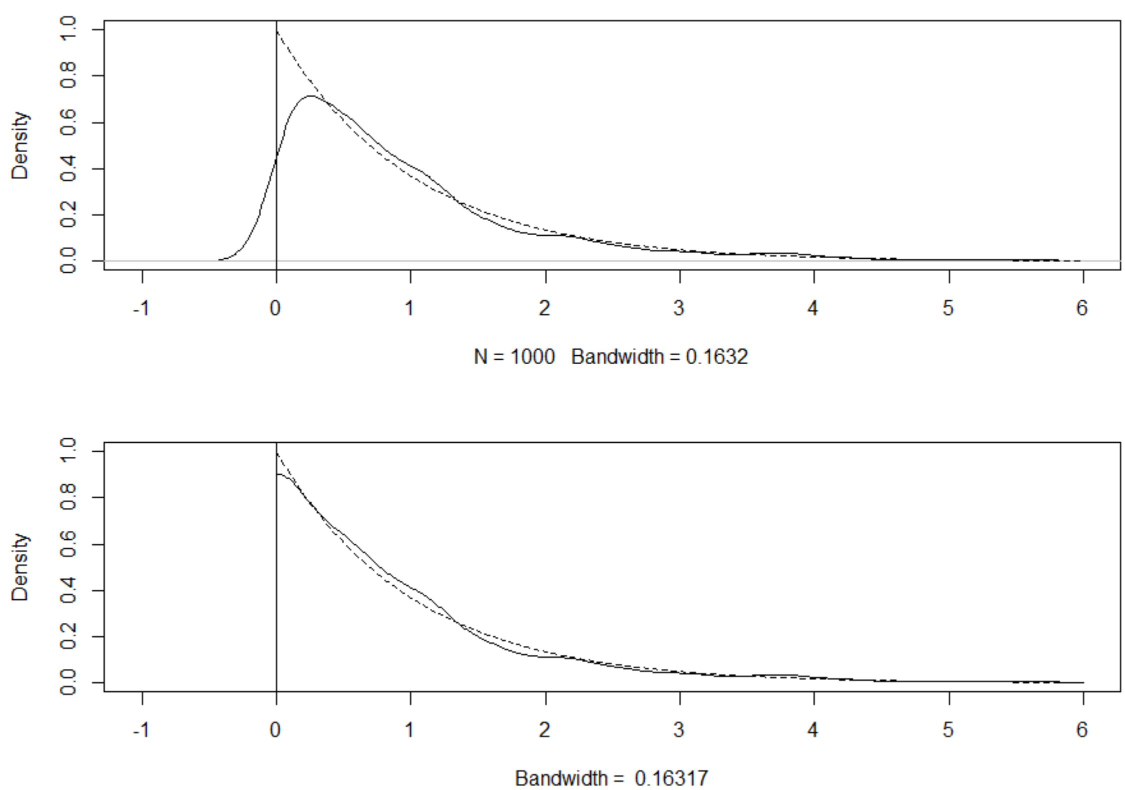


Figure 9 Gaussian kernel density estimates of an exponential density comparison the kernel estimate with true density

Gaussian KDE(solid line) of an exponential density, with true density(dashedline). 2nd plot is the reflection boundary technique.

2.5.1 Reflection Boundary Technique

Generates a class of boundary corrected estimators having properties(local smoothness,nonnegativity).This method performs well when compared with the existing methods for almost all shapes of densities

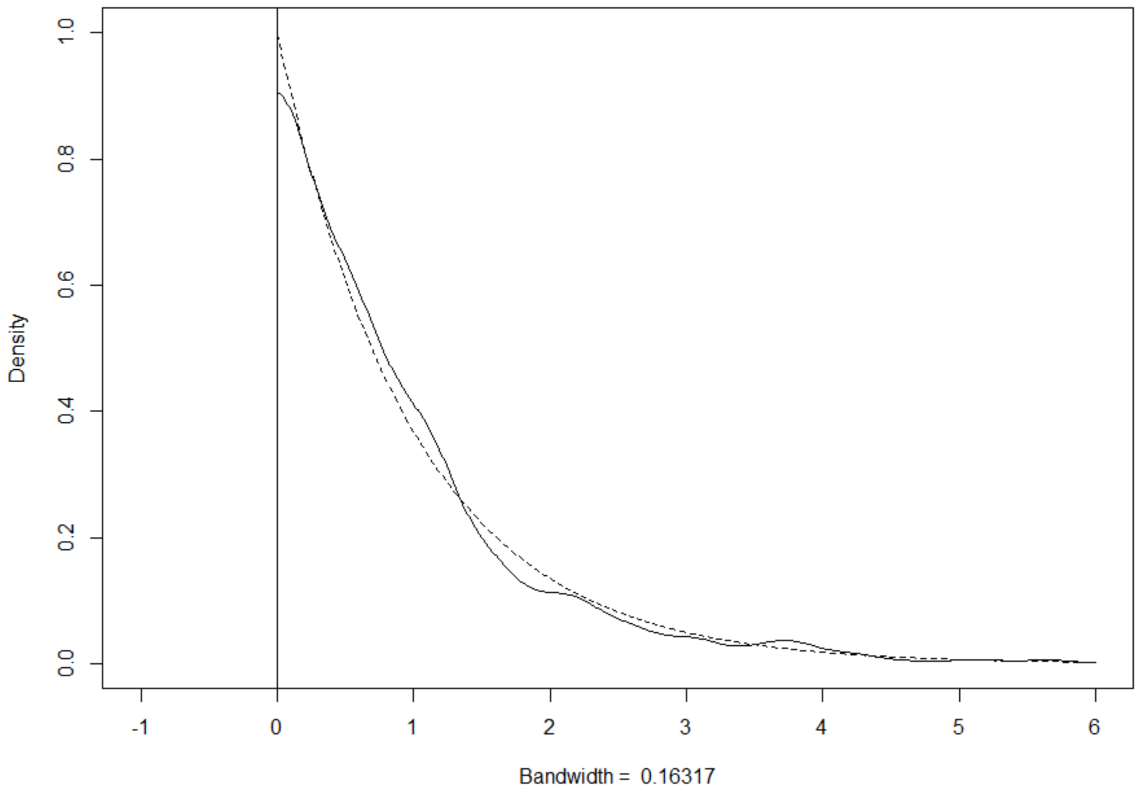


Figure 10 Reflection boundary Technique

3 Bivariate and Multivariate Density Estimations

Bivariate Density Estimations

For a bivariate random sample drawn from density f, KDE is defined by

$f\widehat{(x;H)} = \frac{1}{n}\sum_{i=1}^{n} K(H)\,(x - Xi)$

K(x):symmetric probability density function,H:bandwidth matrix symmetric and positive-definite

K(x) = 1/(2π) * exp(− 1/ 2 x^Tx)

K is not crucial,H is crucial in determining the performance of ˆf The most common parameterizations of the bandwidth matrix are the diagonal and the general or unconstrained hasn’t restrictions on H.H remains positive definite and symmetric

3.1 Bivariate Frequency Polygon

To construct a bivariate density histogram (polygon), necessary to define two-dimensional bins and count the observations in each bin.

Example(Bivariate_frequency_table:bin2d):

The frequencies are computed by constructing a two dimensional contingency table with the marginal breakpoints as the cut points. The return value of bin2d is a list including the table of bin frequencies, vectors of breakpoints, and vectors of midpoints.

```
> bin2d(iris[1:50,1:2])
$call
bin2d(x = iris[1:50, 1:2])

$freq

            (2,2.5] (2.5,3] (3,3.5] (3.5,4] (4,4.5]
  (4.2,4.4]       0       3       1       0       0
  (4.4,4.6]       1       0       3       1       0
  (4.6,4.8]       0       2       5       0       0
  (4.8,5]         0       2       8       2       0
  (5,5.2]         0       0       6       4       1
  (5.2,5.4]       0       0       2       4       0
  (5.4,5.6]       0       0       1       0       1
  (5.6,5.8]       0       0       0       2       1

$breaks1
[1] 4.2 4.4 4.6 4.8 5.0 5.2 5.4 5.6 5.8

$breaks2
[1] 2.0 2.5 3.0 3.5 4.0 4.5

$mids1
[1] 4.3 4.5 4.7 4.9 5.1 5.3 5.5 5.7

$mids2
[1] 2.25 2.75 3.25 3.75 4.25

>
```

Figure 11 Bivariate frequency table

Example(Bivariate_density_polygon)

Bivariate data is displayed in a 3D density polygon, using the bin2d function in the above Example to compute the bivariate frequency table

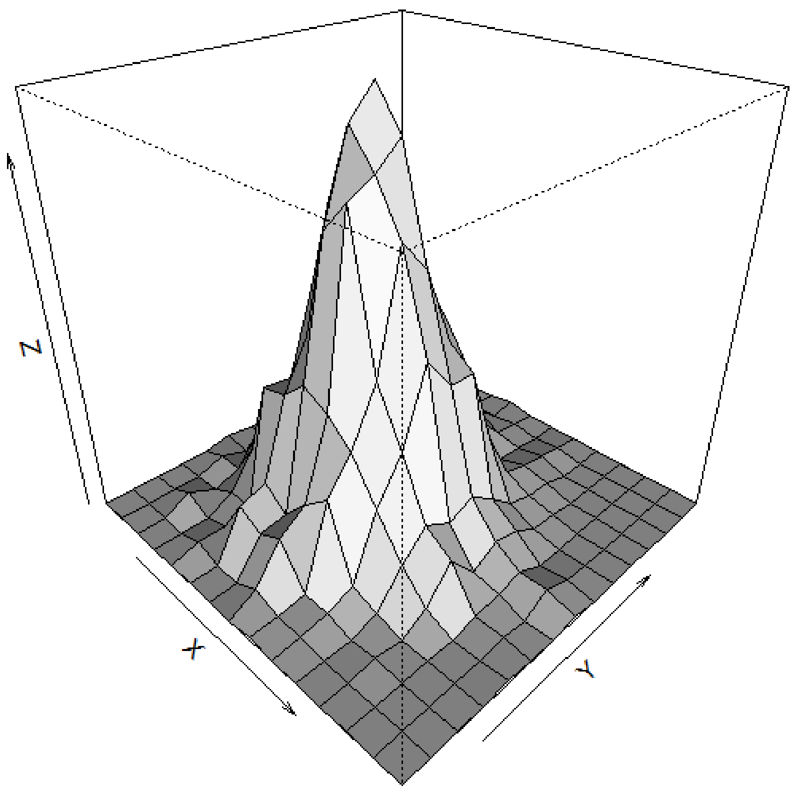


Figure 12 Density polygon of bivariate normal data using normal reference rule (Sturges' Rule) to determine bin widths

3.2 Bivariate ASH

The ASH estimator of density can be extended to multivariate density estimation. Suppose that bivariate data {(x,y)} , have been sorted into an nbin1 by nbin2 array of bins with frequencies ν = (νij) and bin widths

h = (h1,h2) is the number of shifted histograms on each axis used in the estimate. The histograms are shifted in two directions, so that there are m1m2 histogram density estimates to be averaged.

The bivariate ASH estimate of the joint density f(x, y) is

$$\boldsymbol{f}_{\boldsymbol{ASH}}\widehat{(\boldsymbol{x},\boldsymbol{y})} = \frac{\mathbf{1}}{\boldsymbol{m1m2}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{m1}}\sum_{\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{m2}}\boldsymbol{f}_{\boldsymbol{ij}}\widehat{(\boldsymbol{x},\boldsymbol{y})}$$

The bin weights are given by

$\omega_{\boldsymbol{ij}} = \left(\mathbf{1} - \frac{|\boldsymbol{i}|}{\boldsymbol{m1}}\right) * \left(\mathbf{1} - \frac{|\boldsymbol{j}|}{\boldsymbol{m2}}\right)$ $i = 1-m_1,\ldots,m_1-1,\ j = 1-m_2,\ldots,m_2-1.$

Example_(Bivariate_ASH_density_estimate)

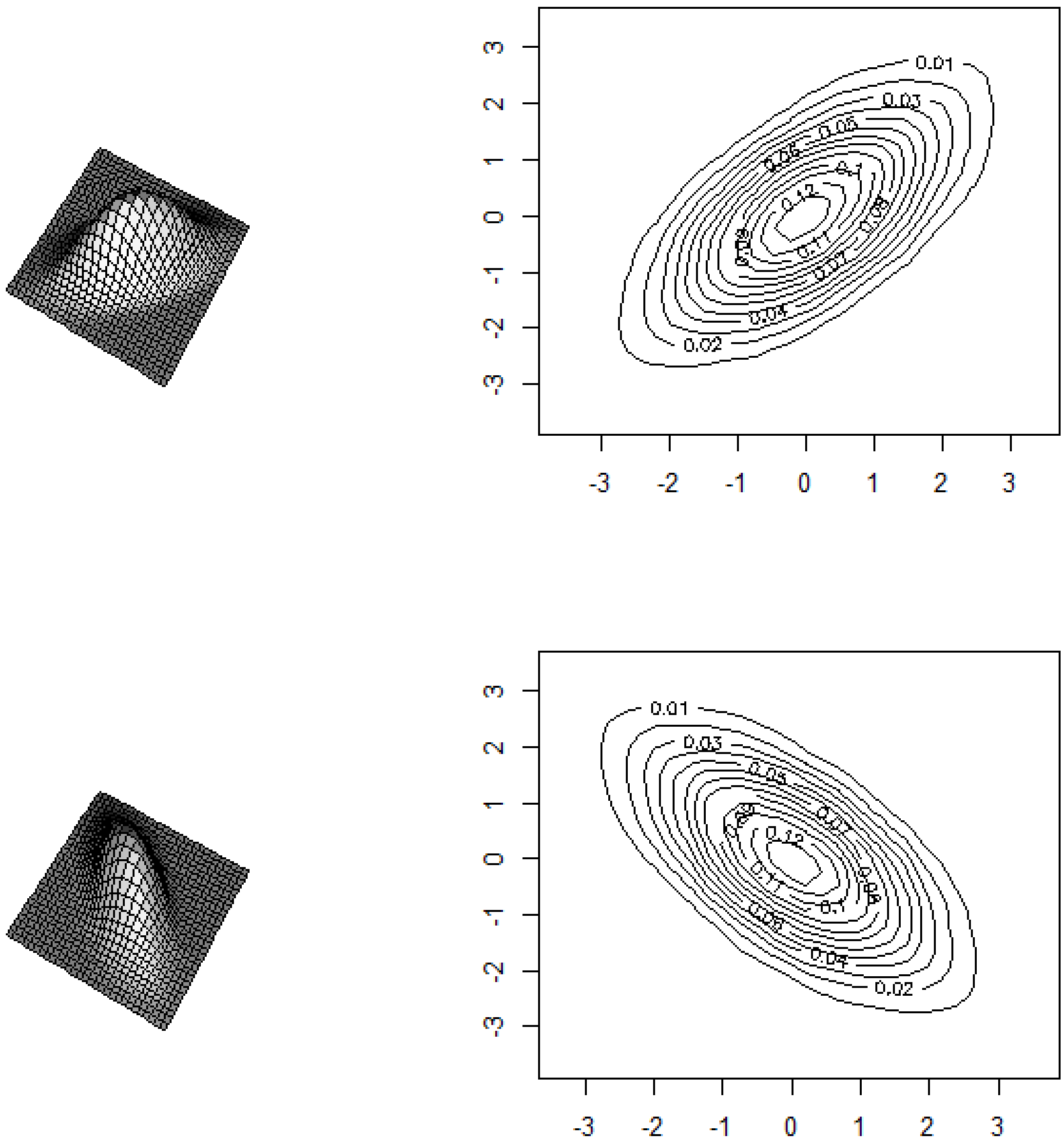


Figure 13 Bivariate ASH density estimates of bivariate normal data

The perspective and contour plots from the ASH estimates are shown in Figures (a) and (b). The variables in the first example have positive correlation $\rho = 0.9$. In the second example, the variables have negative correlation $\rho = -0.9$

3.3 Multidimensional Kernel Methods

X is a random vector in IR^d, and K(X) : $IR^d \rightarrow$ IR is a kernel function, K(X) is a density function on IR^d. n × d matrix(x_{ij}) be sample from the distribution of X. The smoothing parameter is a d-dimensional vector h. If the bandwidth is equal in all dimensions, the multivariate KDE of f(X) with smoothing parameter h_1 is

$$\widehat{fk(x)} = \frac{1}{nh_1^d}\sum_{i=1}^{n} K\,(\frac{X-xi}{h1})$$

xi· i th row of (xij),K(X) symmetric and unimodal density on IR^d The Gaussian kernels have unbounded support.

Example of a kernel with bounded support is the multivariate version of the Epanechnikov kernel,

$K(x)= \frac{1}{2c_d}(d+2)(1-X^TX)I(X^TX<1)$

$Cd=2\pi^{d/2}/(dT(d/2))$) the volume of the d-dimensional unit sphere. d = 1 the constant is c1 = 2 and

K(x) = (3/4)(1−x2) I(|x| < 1), the univariate Epanechnikov kernel(Table_2)

When h1=h2 and the standard Gaussian kernel corresponds to centering identical weight functions like smooth bumps at each sample point and summing the heights of these surfaces to obtain the density estimate at a given point the small bumps will be surfaces (bivariate normal densities) rather than curves. The product KDE of f(X) with smoothing parameter h = (h1,...,hd) is

$$\widehat{f(x)} = \frac{1}{nh1h2 \ldots hd}\sum_{i=1}^{n}\prod_{j=1}^{d} K(\frac{X-xi}{h1})$$

The optimal smoothing parameter for multivariate frequency polygon has $h^*_j = O(n^{-1/(4+d)})$,

AM $ISE^* = O(n^{-4/(4+d)})$,

the optimal smoothing parameter for uncorrelated multivariate normal data

$h^*_j =(\frac{4}{d+2})^{1/(d+2)} \times \sigma i\ n^{-1/(d+4)}$

1/(d+4)converges to 1 ad d tends to infinity

Example(Product_kernel_estimate_of_a_bivariate_normal_mixture)

This example plots the density estimate for a bivariate normal location mixture. The mixture has three components with different mean vectors and identical variance Σ = I2. The mean vectors are

$\boldsymbol{\mu 1} = \begin{bmatrix}\mathbf{0}\\ \mathbf{1}\end{bmatrix}$ $\boldsymbol{\mu 2} = \begin{bmatrix}\mathbf{4}\\ \mathbf{0}\end{bmatrix}$ $\boldsymbol{\mu 3} = \begin{bmatrix}\mathbf{3}\\ \mathbf{-1}\end{bmatrix}$ and the mixing probabilities are p = (0.2, 0.3, 0.5).

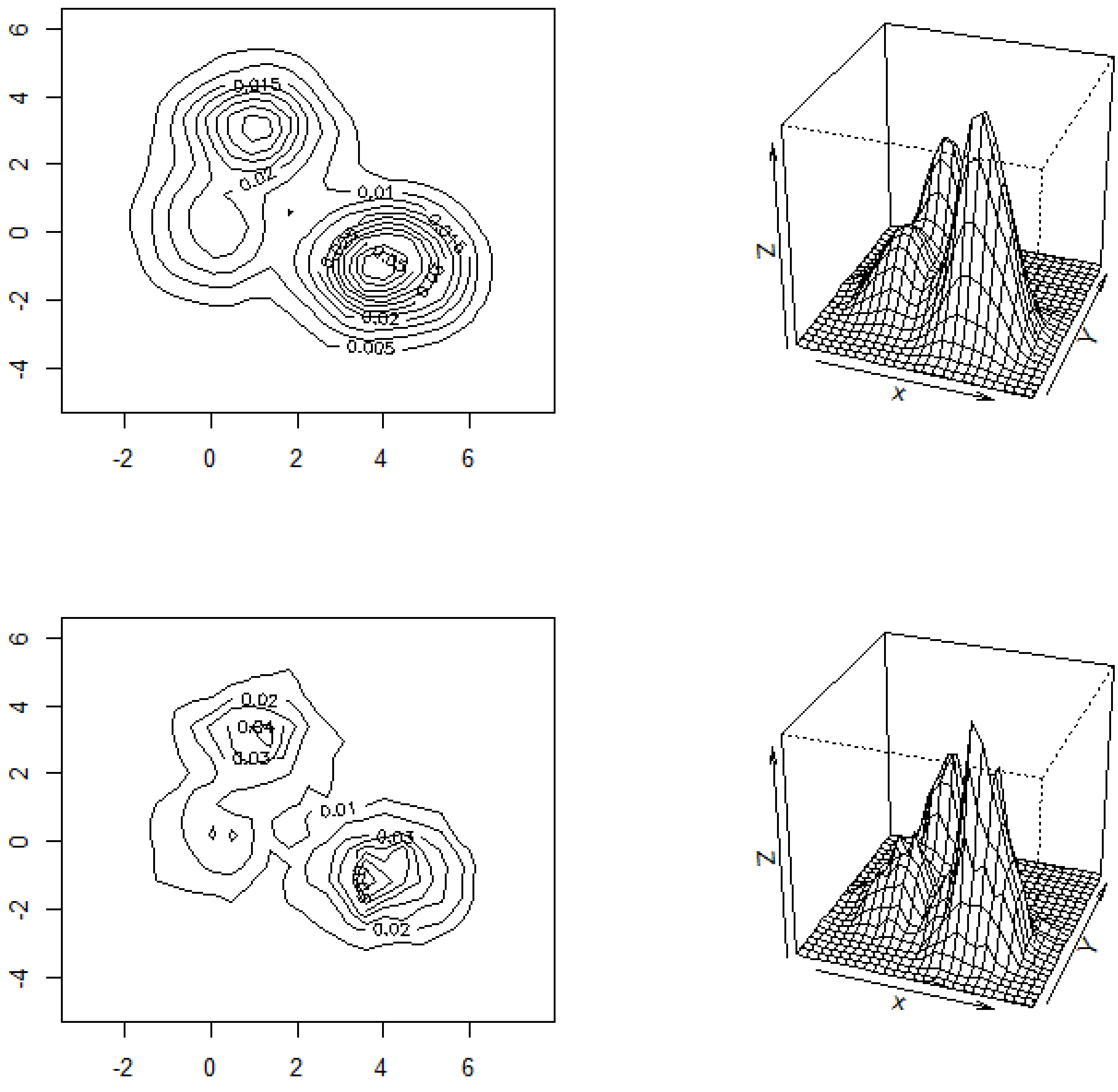


Figure 14 Product kernel estimates of bivariate normal mixture data

4 Other Methods of Density Estimation

4.1 Fourier Expansion Approach

The estimation of the density f of X is to represent f by its **Fourier expansion** and estimate **the coefficients** from sample, The estimator is not useful,it will tend to a sum of delta functions that place probability mass at the individual observations. This is resolved by smoothing to obtain a more useful density estimator, and generalizing to densities with unbounded support

4.2 Adaptive Kernel Methods

Adaptive KDE is an efficient estimator when the density has long tail or multi-mode. Varying bandwidths at each observation point used by adapting a fixed bandwidth for data

4.3 K Near Neighbor Estimates

The KNN method estimates the density value at point x based on the distance between x and its kth nearest neighbor. A large KNN distance indicates that the density is usually small, and vice versa

4.4 Penalized Likelihood Methods

Assume a bounded domain X so that the uniform density is proper. f(x) = e^{nx} / $\int x$ e^n(x) the logistic density transform $\widehat{f(x)}$ can be conducted through the minimization of a penalized likelihood functional,

$-\frac{\mathbf{1}}{\boldsymbol{n}}\sum\{\eta(\boldsymbol{xi}) - \boldsymbol{log} \int \boldsymbol{e}^{\boldsymbol{n}(\boldsymbol{x}\}} + \boldsymbol{\lambda}/\mathbf{2}\boldsymbol{J}(\eta)\}$

4.5 Parametric Density Estimations

Parametric density estimation assumes that the data are from a known family of distributions, such as the normal, lognormal, exponential, and Weibull and estimating the parameters from the sample.

5 References

1. **https://onlinelibrary.wiley.com/doi/abs/10.1002/cjs.5550330403**

2. **http://students.aiu.edu/submissions/profiles/resources/onlineBook/z2n4t4_Multivariate_Density_Estimation.pdf**

3. **http://yaroslavvb.com/papers/scott-multivariate.pdf**

4. **https://www.jstor.org/stable/2290332**

5. **https://faculty.engineering.ucdavis.edu/lai/wp-content/uploads/sites/38/2021/06/knndensityv8.pdf**

6. **https://www.jstor.org/stable/24307125**

7. **https://www3.stat.sinica.edu.tw/statistica/oldpdf/a13n314.pdf**

8. **http://faculty.washington.edu/yenchic/18W_425/Lec8_parametric.pdf**

6 Appendix:

Google drive link:
https://drive.google.com/drive/folders/1QW21ktMpYuE_T7M_QcUnX8_IFV4a1L7c?usp=sharing

- INDEX Number : s15680